EXHIBIT H

Journal of Consulting Psychology
1965, Vol. 29, No. 3, 261–265

OVERCONFIDENCE IN CASE-STUDY JUDGMENTS[1]

STUART OSKAMP

Claremont Graduate School

This study investigated whether psychologists' confidence in their clinical decisions is really justified. It was hypothesized that as psychologists study information about a case (a) their confidence about the case increases markedly and steadily but (b) the accuracy of their conclusions about the case quickly reaches a ceiling. 32 judges, including 8 clinical psychologists, read background information about a published case, divided into 4 sections. After reading each section of the case, judges answered a set of 25 questions involving personality judgments about the case. Results strongly supported the hypotheses. Accuracy did not increase significantly with increasing information, but confidence increased steadily and significantly. All judges except 2 became overconfident, most of them markedly so. Clearly, increasing feelings of confidence are not a sure sign of increasing predictive accuracy about a case.

It is a common phenomenon of clinical practice that as a psychologist accumulates case-study material about another human being, he comes to think that he knows that person pretty well. Consequently, sooner or later in the information-gathering process, the psychologist becomes confident enough to make diagnostic conclusions, describe the client's main dynamics, and perhaps even venture to predict his future behavior. Though the psychologist's conclusions may remain tentative, his increase in confidence from the time of first approaching the case to the time of writing his report is usually very marked.

This study investigated whether that increase in confidence is justified by a corresponding increase in accuracy of conclusions. Though the psychologist's confidence in his conclusions has often been mentioned as an important subject of scientific inquiry (Meehl, 1957), it has only rarely been studied intensively. Furthermore, when it has been studied, rather surprising findings have often resulted. For instance, Goldberg (1959) and Oskamp (1962) have shown that the diagnostic confidence of experienced psychologists is *less* than that of less experienced persons. The same studies and many others have also shown that professional psychologists are no better interpersonal judges, and sometimes are worse ones, than are untrained individuals (Taft, 1955).

Another rarely studied factor, which may provide a good index of the expertness of a judge, is the relationship between his level of confidence and his level of accuracy. This measure shows, for instance, whether the judge is overconfident or underconfident in making his decisions. On this measure, which may be termed appropriateness of confidence, experienced judges have been found to be far superior to inexperienced ones (Oskamp, 1962).

A number of studies (Hamlin, 1954; Hathaway, 1956; Kostlan, 1954; Soskin, 1954; Winch & More, 1956) have investigated the effects on clinical judgment of differing amounts of stimulus information. In the present experiment this factor was studied by giving each judge four sets of cumulatively increasing amounts of information as the basis for making his decisions, thus simulating the gradual buildup of information as a psychologist works his way through a typical case.

The hypotheses of the study were as follows:

1. Beyond some early point in the information-gathering process, predictive accuracy reaches a ceiling.

2. Nevertheless, confidence in one's decisions continues to climb steadily as more information is obtained.

3. Thus, toward the end of the information-gathering process, most judges are overconfident about their judgments.

[1] Revision of a paper presented at the Western Psychological Association annual meeting, April 18, 1964. Thanks are due to Stanley Lunde for his help in constructing the case-study test.

PROCEDURE

Since it was desired to simulate the usual clinical situation as closely as possible, an actual case study was chosen as the information to be given to the judges. The case finally chosen was selected because of its extensiveness, its description of many pertinent life incidents, and the fact that it involved a relatively normal individual (i.e. a case of adolescent maladjustment who had never been psychiatrically hospitalized). It was the case of Joseph Kidd, reported by White (1952) in his book, *Lives in Progress*.[2]

Historical background material from this case was summarized and organized into chronological sets of information which were presented to the judges in four successive stages. Stage 1 contained only the following brief demographic information about the case, in order to test for the "psychological chance" level of predictive accuracy (Patterson, 1955):

> Joseph Kidd (a pseudonym) is a 29 year old man. He is white, unmarried, and a veteran of World War II. He is a college graduate, and works as a business assistant in a floral decorating studio.

Stage 2 added 1½ single-spaced typed pages of material about Kidd's childhood, through age 12. Stage 3 (2 pages) covered his high school and college years, and Stage 4 (1⅓ pages) covered his army service and later activities up to age 29.

Case-Study Test

In order to have a basis for determining the accuracy of the judges, a multiple-choice case-study test was constructed, using a method similar to that of Soskin (1954). Items dealt with Kidd's customary behavior patterns, attitudes, interests, and typical reactions to actual life events. Examples of some of these items are given in Table 1.

Items were constructed only where there was fairly objective criterion information presented in the case, either factual data or well-documented conclusions. The four incorrect alternatives for each item were made up with the help of sentence-completion responses to the item stems by psychology graduate students. They were constructed in such a way as to be clearly wrong, based on the published case material, but to be otherwise convincing and "seductive" alternatives. None of the items had their answers contained in the summarized case material; instead, judges were expected to follow the usual procedure in clinical judgment (McArthur, 1954) by forming a personality picture of Kidd from the material presented and then predicting his attitudes and typical actions from their personality picture of him.

[2] Use of this case had the disadvantage that a few judges remembered reading this material at some time during their training, but all but one reported that their earlier contact did not help them at all in the present study. Since their accuracy scores corroborated this impression, their results were retained in the data analysis.

Judges

Judges were drawn from three groups with varying amounts of psychological experience: (*a*) 8 clinical psychologists employed by a California state hospital, all of whom had several years of clinical experience, and 5 of whom had doctor's degrees;[3] (*b*) 18 psychology graduate students;[4] and (*c*) 6 advanced undergraduate students in a class in personality. None of the judges was in any way familiar with the hypotheses of the study.

Judges took part in the experiment in small groups ranging from four to nine in size, but each worked at his own individual pace with his own sheaf of materials. After reading each stage of the case, the judge answered all 25 questions of the case-study test before going on to read the next stage. In addition to answering the questions, the judge also indicated on each item how confident he was that his answer was correct.

Confidence Judgments

The confidence judgments were made using a scale devised by Adams (1957) which defines confidence in terms of expected percentage of correct decisions. Since there were five alternatives for each test item, the scale began at 20% (representing a completely chance level of confidence) and extended to 100% (indicating absolute certainty of correctness). In addition to providing a clearly understood objective meaning for confidence, this scale has the great advantage of allowing a direct comparison between the level of accuracy and the level of confidence. Thus, for example, if a judge got 28% of the items correct and had an average confidence level of 43%, he could clearly be said to be overconfident.

RESULTS

This judgment task proved to be a very difficult one, at least with the amount of case material provided. No judge ever reached 50% accuracy, and the average final accuracy was less than 28%, where chance was 20% (a non-significant difference). However, this low level of accuracy serves to provide an even more dramatic test of the hypotheses of the study.

A preliminary analysis was carried out to compare the scores of the three groups of judges, though no hypotheses had been formulated about their relative performance. These results clearly indicated that there were no significant differences among the three

[4] About half of these graduate students had had some clinical or counseling experience, and one or two may possibly have been equivalent to the clinical psychologists in level of psychological experience.

[3] One additional clinical psychologist was tested, but results had to be discarded due to failure to understand and follow the instructions. This problem did not occur with any of the students.

TABLE 1

SAMPLE ITEMS FROM THE CASE-STUDY TEST

5. During college, when Kidd was in a familiar and congenial social situation, he often:

   a. Tried to direct the group and impose his wishes on it.
   b. Stayed aloof and withdrawn from the group.
   c. Was quite unconcerned about how people reacted to him.
   d. Took an active part in the group but in a quiet and modest way.
   e. Acted the clown and showed off.[a]

10. Later during his Army service, as an officer and detachment commander, Kidd's attitude toward handing out punishment was:

   a. He was very disturbed by it because he preferred to be on the same level as other men, not over them.[a]
   b. He disliked it because he could never make a decision as to what to do.
   c. He avoided it as completely as possible because he felt that it was wrong to punish men no matter what they had done.
   d. He was happy because it gave him a chance to be in control of a situation and to be looked up to.
   e. He took a sadistic delight in disciplining others to make up for the times he had been punished.

15. Kidd's present attitude toward his mother is one of:

   a. Love and respect for her ideals.
   b. Affectionate tolerance for her foibles.
   c. Combined respect and resentment.[a]
   d. Rejection of her and all her beliefs.
   e. Dutiful but perfunctory affection.

20. In conversations with men, Kidd:

   a. Prefers to get them to talk about their work or experiences.[a]
   b. Likes to do most of the talking about subjects with which he is familiar.
   c. Prefers to debate with them about religion or their philosophy of life.
   d. Likes to brag about his Army days or college exploits.
   e. Confines his discussion mainly to sports, sex, and dirty jokes.

25. Kidd's attitude toward his life as a business assistant is shown by his recent decision to:

   a. Stay in his present position for at least a few more years.
   b. Expand the business by building another shop in a nearby town.
   c. Leave his job and open up his own flower shop.
   d. Make job applications to several larger companies in fields similar to his present line of work.
   e. Strike out on his own and find a different kind of job.[a]

[a] Correct answer.

groups of judges either in accuracy, in confidence, or in total number of changed answers. The Stage 4 confidence scores were consistent with previous studies (Goldberg, 1959; Oskamp, 1962) in showing the more experienced judges to be *less* confident than the less experienced judges, but in this study these results did not approach significance.

The main results of the study are shown in Table 2, where the successive columns show the judges' mean scores as they received successively greater amounts of information. As a result of the previous statistical tests, results for all 32 judges are combined in this table.

The first line of Table 2 shows that the fluctuation in accuracy over the four stages of the case was significant. However, a Duncan multiple-range test (Edwards, 1960, p. 136) showed that this significance was due primarily to the drop in accuracy at Stage 2. Comparing Stage 1 accuracy with Stage 4 accuracy showed no significant change ($t = 1.13$, $df = 31$). Thus, the first hypothesis concerning a ceiling on accuracy was not only supported, but in this experiment there was no significant increase in accuracy at all with increasing information!

Hypothesis 2 is tested in the second line of

Table 2. There we see, as predicted, a striking and extremely significant rise in confidence from 33% at Stage 1 to 53% at Stage 4.

Finally, results of Hypothesis 3 are indicated by a comparison of the first and second lines of the table. At Stage 1 the average amount of overconfidence was 7 points; at Stage 4 it was 25 points, a difference significant far beyond the .001 level ($t = 5.14$, $df = 31$).

Sometimes group means may be significant but misleading because they may conceal individual subjects who perform contrary to prediction. That this was not the case here is clearly shown by the following figures for individual judges. Of the 32 judges, 14 increased in accuracy from Stage 1 to Stage 4, while 6 remained the same, and 12 decreased—a completely random result. By contrast, all judges except 2 increased in confidence, and most increased markedly.[5] At Stage 1 almost half of the judges (13 out of 32) were not overconfident; by Stage 4 only 2 remained underconfident—a highly significant change ($\chi^2 = 9.1$, $p < .01$).

Another interesting result of the study is contained in the last line of Table 2, which shows the average number of items on which the judges changed their answers at each stage of the case. This measure shows that as more information was presented, the number of changed answers *decreased* markedly and significantly. This finding suggests that the judges may frequently have formed stereotype conclusions rather firmly from the first fragmentary information and then been reluctant to change their conclusions as they received new information. At any rate, the final stage of information seems to have served mainly to confirm the judges' previous impressions rather than causing them to revamp their whole personality picture of Kidd.

[5] One of the two judges who decreased in confidence, an undergraduate, later stated that he would normally have increased in confidence, but he had just been engaged in a computer research project in which the computer had repeatedly given incorrect results, to the point where he had completely lost his confidence even in computers.

DISCUSSION

Careless generalization of these findings must certainly be avoided. There are three main factors about this study which might possibly limit the generality of the results. (*a*) The case may not be similar to the ones with which most psychologists are used to working. (*b*) The test items may not represent the sorts of behaviors which psychologists are used to predicting. (*c*) The judges may not have been good representatives of psychological decision makers. In answer to these possible objections it should be pointed out that the case, the test items, and the clinical judges were all chosen with the intention of approximating as closely as possible the situations found in actual psychological practice.

Even if these possible objections were to be granted though, some clear-cut conclusions can be drawn. Regardless of whether the task seemed strange or the case materials atypical, the judges' confidence ratings show that *they became convinced of their own increasing understanding of the case.* As they received more information, their confidence soared. Furthermore, their certainty about their own decisions became entirely out of proportion to the actual correctness of those decisions.

Thus, though this result may not hold for every psychologist and every type of decision,

TABLE 2

PERFORMANCE OF 32 JUDGES ON THE 25-ITEM CASE-STUDY TEST

Measure	Mean Score				F	p
	Stage 1	Stage 2	Stage 3	Stage 4		
Accuracy (%)	26.0	23.0	28.4	27.8	5.02	.01
Confidence (%)	33.2	39.2	46.0	52.8	36.06	.001
Number of changed answers	—	13.2	11.4	8.1	21.56	.001

it can clearly be concluded that a psychologist's increasing feelings of confidence as he works through a case are *not* a sure sign of increasing accuracy for his conclusions. So-called clinical validation, based on the personal feelings of confidence of the clinician, is not adequate evidence for the validity of clinical judgment in diagnosing or predicting human behavior.

REFERENCES

ADAMS, J. K. A confidence scale defined in terms of expected percentages. *American Journal of Psychology,* 1957, **70,** 432–436.

EDWARDS, A. E. *Experimental design in psychological research.* New York: Holt, Rinehart, & Winston, 1960.

GOLDBERG, L. R. The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt Test. *Journal of Consulting Psychology,* 1959, **23,** 25–33.

HAMLIN, R. M. The clinician as judge: Implications of a series of studies. *Journal of Consulting Psychology,* 1954, **18,** 233–238.

HATHAWAY, S. R. Clinical intuition and inferential accuracy. *Journal of Personality,* 1956, **24,** 223–250.

KOSTLAN, A. A method for the empirical study of psychodiagnosis. *Journal of Consulting Psychology,* 1954, **18,** 83–88.

MCARTHUR, C. Analyzing the clinical process. *Journal of Counseling Psychology,* 1954, **1,** 203–207.

MEEHL, P. E. When shall we use our heads instead of the formula? *Journal of Counseling Psychology,* 1957, **4,** 268–273.

OSKAMP, S. The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs,* 1962, **76** (28, Whole No. 547).

PATTERSON, C. H. Diagnostic accuracy or diagnostic stereotypy? *Journal of Consulting Psychology,* 1955, **19,** 483–485.

SOSKIN, W. F. Bias in postdiction from projective tests. *Journal of Abnormal and Social Psychology,* 1954, **49,** 69–74.

TAFT, R. The ability to judge people. *Psychological Bulletin,* 1955, **52,** 1–23.

WHITE, R. W. *Lives in progress: A study of the natural growth of personality.* New York: Dryden Press, 1952.

WINCH, R. F., & MORE, D. M. Does TAT add information to interviews? Statistical analysis of the increment. *Journal of Clinical Psychology,* 1956, **12,** 316–321.

(Received July 17, 1964)